Published online 10 May 2010

Nucleic Acids Research, 2010, Vol. 38, No. 17 5623–5633
doi:10.1093/nar/gkq343

Mathematical modelling of whole chromosome replication

Alessandro P. S. de Moura[1], Renata Retkute[2], Michelle Hawkins[2] and Conrad A. Nieduszynski[2,*]

[1]Department of Physics, University of Aberdeen, Aberdeen AB24 3UE and [2]School of Biology, University of Nottingham, Nottingham NG7 2UH, UK

Received March 5, 2010; Revised April 12, 2010; Accepted April 19, 2010

ABSTRACT

All chromosomes must be completely replicated prior to cell division, a requirement that demands the activation of a sufficient number of appropriately distributed DNA replication origins. Here we investigate how the activity of multiple origins on each chromosome is coordinated to ensure successful replication. We present a stochastic model for whole chromosome replication where the dynamics are based upon the parameters of individual origins. Using this model we demonstrate that mean replication time at any given chromosome position is determined collectively by the parameters of all origins. Combining parameter estimation with extensive simulations we show that there is a range of model parameters consistent with mean replication data, emphasising the need for caution in interpreting such data. In contrast, the replicated-fraction at time points through S phase contains more information than mean replication time data and allowed us to use our model to uniquely estimate many origin parameters. These estimated parameters enable us to make a number of predictions that showed agreement with independent experimental data, confirming that our model has predictive power. In summary, we demonstrate that a stochastic model can recapitulate experimental observations, including those that might be interpreted as deterministic such as ordered origin activation times.

INTRODUCTION

Accurate and complete replication of the genome is crucial for life. All chromosomes in eukaryotic cells must be replicated exactly once and then segregated to daughter cells to ensure genetic integrity. Mistakes in chromosome inheritance play a major role in human diseases such as cancers and congenital disorders. During S phase, replication of eukaryotic genomes is initiated at multiple discrete chromosomal sites called replication origins. The yeast genome contains hundreds of origins (1) and metazoan genomes contain thousands (2).

Completion of genome replication prior to cell division requires the activation of a sufficient number of appropriately distributed origins. Not all origins are activated during a cell cycle. The presence of dormant origins that are passively replicated by forks from a neighbouring origin may provide a means to overcome replication impediments (3,4). Therefore, it is necessary to investigate how the activity of multiple origins on each chromosome is coordinated to ensure successful replication.

Regulation of DNA replication is divided into two temporally distinct stages: the establishment of activation-competent origins and their subsequent activation. When cyclin-dependent kinase (CDK) activity is low (late mitosis and G1 phase) origin competence is established in a step called 'licensing'. Licensing involves assembly of a series of proteins (Orc1-6, Cdc6, Cdt1 and Mcm2-7) at the origin to form the pre-replication complex (pre-RC). Pre-RC assembly marks origins as competent for initiation. Increasing CDK activity, as cells pass from G1 to S phase, inhibits further pre-RC assembly, thus ensuring only one round of DNA replication per cell cycle. In S phase the activity of two kinases, S phase CDK and Cdc7, initiates DNA replication. Upon origin activation, departure of the elongating replication forks, including Mcm2-7, results in origin inactivation. In addition, if a fork initiated at a neighbouring origin 'passively' replicates an origin, the licensing factors are displaced and the origin is inactivated (5–7).

Eukaryotic replication origins are best understood in *Saccharomyces cerevisiae*, where specific origin sequences have been isolated (8). Every origin contains an essential sequence element (called the ACS) that is the binding site for Orc1-6 (9,10). A variety of complementary approaches have mapped budding yeast origins genome-wide.

*To whom correspondence should be addressed. Tel: +44 115 823 0352; Fax: +44 115 823 0338; Email: conrad.nieduszynski@nottingham.ac.uk

We used phylogenetic foot-printing to identify ACS sites (11), while others have used microarrays to determine Orc- (12) and Mcm-binding sites (13). A third method has involved identifying origins as the earliest replicating sites (14–16). These studies provide a platform for understanding chromosome and genome replication.

S. cerevisiae is frequently used as a paradigm for investigating eukaryotic cellular processes and for mathematical modelling of biological pathways (17). Our understanding of *S. cerevisiae* DNA replication has advanced to the point where we can define the parameters that underpin the entire replication system. Values are known for many of these parameters. Measurements have also been made of replication system outputs; including the mean time during S phase at which each region of the genome replicates (15,16). *S. cerevisiae* is therefore an ideal organism in which to formulate and validate a mathematical model of chromosome replication. Such a model should allow quantitative predictions about chromosome replication, including the times and efficiencies of origin activation. Despite this wealth of data there have been few attempts to mathematically model *S. cerevisiae* chromosome replication (18,19). Previous mathematical models of eukaryotic chromosome replication have focussed on understanding how origin distribution ensures complete chromosome replication—the 'random completion problem' (20–24). *S. cerevisiae* replication origins have well-defined chromosomal locations effectively removing the random completion problem. This might be a consequence of the fact that *S. cerevisiae* has several small chromosomes (four chromosomes <500 kb) that could be lost if there were large random regions with no origins.

The complexity of chromosome replication makes a mathematical model important to aid understanding of the intricate dynamics. Approaches based purely upon qualitative reasoning cannot determine how measured quantities, such as replication times, relate to the fundamental parameters of the system, such as the origin positions and competences. Here we present a stochastic model for whole chromosome replication, which we implement in the form of numerical algorithms and explore through computer simulations. We estimate origin parameters from experimental replication time course data. Using these estimated parameters we make a series of testable predictions, including the efficiency of origin usage, and show that these predictions agree with independent experimental data, thus validating our model.

MATERIALS AND METHODS

Stochastic simulation of chromosome replication

We use a Monte Carlo simulation method to generate outputs such as replication timing (T_{rep}) curves and origin efficiencies from the model. Making the assumption that different chromosomes replicate independently, our numerical procedures focus on a single chromosome. A virtual population of N chromosomes is first generated, with each member having origin activation times taken from a Gaussian probability distribution with the given mean (T_i) and variance (σ_i) of each origin; typical values of N used in our simulations varied from 2000 to 1 000 000. Origins may fail in each chromosome in the population with a probability $1-p_i$ (where p_i is their competence), in which case they are inactive on the corresponding chromosomes. Then the replication time of a given position in the chromosome for a particular member of the population is found by evaluating the minimum of the quantities $\tau_i = t_i + |x - x_i|/v$, where the index i runs over all activating origins on that chromosome. Here, x is the position the replication time is calculated for; x_i is the position of the ith origin; t_i is the activation time of that origin in this particular chromosome; and v is the fork velocity. τ_i is the time it would take for a fork originated at origin i to reach position x. Taking the minimum value of τ_i results in chromosome position (x) being replicated by the first fork to arrive. This procedure efficiently gives the replication time of a single 'cell' in our virtual population; repeating this for each of the N members of the population yields an ensemble of results from which population averages can be taken to obtain data such as T_{rep} curves, which can be compared directly with experimental data.

Origin efficiency for a given origin i is calculated using the same procedure described above, with x being x_i, the position of the origin itself; and keeping track in each member of the virtual population of whether the position $x = x_i$ was replicated by the ith origin, or by another origin. After this is done for the whole population, we count the number n_i of cells in which the ith origin was replicated by forks initiated at other origins, and the efficiency E is given by $E = 1 - n_i/N$. The simulation scheme explained above was implemented as a collection of C programs and a library (available upon request).

Sensitivity analysis

The robustness of our model's predictions was ascertained by performing simple sampling sensitivity analysis. We estimated the sensitivity of important model outputs (predicted origin efficiencies) to small random variations of the model's parameters. Small changes in the parameters lead only to small variations in the predictions. For example, restricting the parameters to be chosen in a window of $\pm 5\%$ around the values of the base parameter set (Supplementary Table S1, row 10), we found that on average the predicted origin efficiencies vary within a window of $\sim$4% of their original values; the largest variation in predicted origin efficiency that we observed was only 16%. This confirms that the model is robust against small parameter perturbations, and that no 'fine-tuning' of parameters takes place.

Parameter estimation

After testing various parameter estimation methods, we found that genetic algorithms were the most effective for our system. The fitness (or score) of a given parameter set is defined by the sum of the square of differences between experimental and simulated data (T_{rep} profiles or unsmoothed replicated fractions at different time points).

This means that a simulation must be run every time the fitness needs to be evaluated by the genetic algorithm. Simulated data used a population (N) of 2000. We used a version of the genetic algorithm that replaced 20% of the parameter sets in each generation, using tournament selection, with a crossing-over probability of 80%. For each generation of the Genetic Algorithm 200 parameter sets were evaluated. We allowed each run of the genetic algorithm to continue until either the number of generations hit a prescribed maximum of 1000, or the score showed no improvement over the past 100 generations. We used the existing open-source implementation PGAPack (http://ftp.mcs.anl.gov/pub/pgapack/).

RESULTS

The model

The main factors determining the dynamics of chromosome replication are the properties of the replication origins (Figure 1). To develop a generic model that can be applied to a range of experimental systems and organisms we have defined four origin parameters, which describe each origin i in a given chromosome:

- chromosomal position x_i
- competence p_i: the fraction of cells in which an origin is biochemically competent to activate in S phase
- mean activation time T_i during S phase
- the width of the activation time probability distribution σ_i.

A final system parameter is the replication fork velocity v. This model includes two stochastic components that give rise to differences between cells: origin competence and the distribution of origin activation times. Consistent with this, a recent single molecule study of *S. cerevisiae* chromosome VI replication indicates that there is significant stochasticity in replication origin activation (25). Differences in origin competence could result from origins assembling pre-RCs with differing proficiency (26–28). Although origins are reported to have characteristic activation times, these represent the mean of a distribution due to the stochastic nature of biological systems. The interplay between the above parameters determines which origins will be used in a particular cell cycle and therefore the fraction of cells in which any given origin is active, that is the origin efficiency (Supplementary Data 1 and Supplementary Figure S1).

Simulating the replication of a virtual chromosome

We have numerically implemented our model in the form of a computer simulator. In these simulations we have made four assumptions (discussed further in Supplementary Data 2):

- activation time probability distribution of every origin is Gaussian;
- constant fork velocity irrespective of the direction and chromosomal location of the fork;
- perfect cell-cycle synchrony;

Parameters (illustrated below)
Origin competence is the fraction of cells in which an origin is biochemically competent to activate (p).
Origin activation time is the time during S phase at which the origin activates. We have described this by a mean activation time (T) and the width of the activation time probability distribution (σ).

System Outputs
Origin efficiency describes the fraction of cells in which the origin activates during S phase.
Replication timing (T_{rep}) profiles show the time at which each chromosomal coordinate is replicated in half the cycling population.

Concepts
Replication timing programme refers to the defined order in which chromosomal regions are observed to replicate.
Passive replication of an origin arises when replication forks from a neighbouring origin replicate through and inactivate an origin.
Dormant replication origins are competent origins that do not activate due to passive replication by forks from neighbouring origins.
Identifiability is the extent to which it is possible to uniquely estimate parameters from the experimental data.

Relationships
Dormant Origins + Active Origins = Competent Origins
Competent Origins + Non-Competent Origins = All Origins
Origin Efficiency = Origin competence X the probability that the origin is not passively replicated

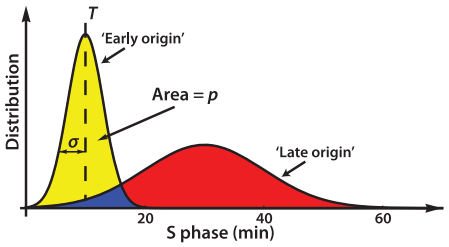

Figure 1. Defining terms.

- no correlation between the parameters of one origin and the parameters of other origins.

To test our model we have used a hypothetical 100 kb virtual chromosome with three origins. All three origins have identical activation time distributions, but one origin (*ORI3*) has a lower competence (Figure 2A). The program then simulates replication of the virtual chromosome in a population of cells and predicts a number of system outputs. These include: replication dynamics of individual chromosomes within the population; the proportion of cells in which each origin activates (origin efficiency); the proportion of replication forks moving in each direction for every chromosomal coordinate; the proportion of every chromosomal coordinate that is replicated at various times through S phase; the distribution of replication termination events; and the time during S phase at which the DNA at a given chromosomal coordinate has been replicated in half the cells (T_{rep}) (Figure 2 and Supplementary Figure S2). By convention T_{rep} values are plotted with time running down the y-axis, so that early replicating sequences appear as peaks and late replicating sequences as valleys (Figure 2C).

The replication of individual chromosomes within the population (Figure 2B and Supplementary Figure S2A) illustrates how all four origin parameters (and the fork velocity) contribute to varied origin usage and a range

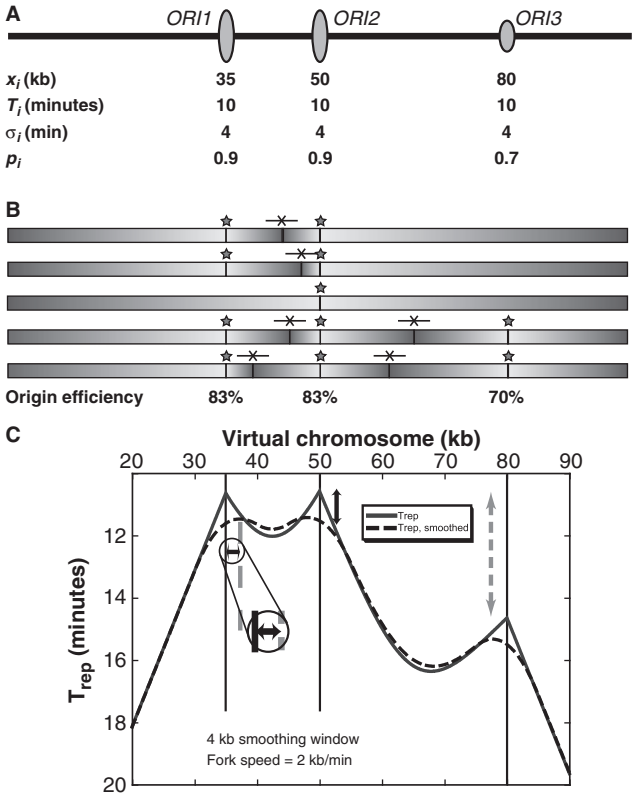

Figure 2. A virtual chromosome replicated by three origins. (**A**) Schematic representation of the virtual chromosome indicating the location of the origins (*ORI*) and their parameters. (**B**) Replication dynamics for example individual chromosomes within the population (the light to dark gradient indicates the direction of replication forks; black bars indicate sites of replication initiation and termination; stars indicate origins that have activated and crosses indicate termination events; see also Supplementary Figure S2A). (**C**) Simulated mean T_{rep} for the virtual chromosome, unsmoothed (solid grey line) and smoothed (dashed black line). The black arrows indicate how peaks shift as a consequence of smoothing. The dashed grey arrow highlights the difference in peak height that results from *ORI3* having a lower competence than the other origins.

of termination sites. The distance between *ORI2* and *ORI3* (30 kb), the fork velocity (2 kb/min), the difference in mean activation time (0 min) and the width of the activation time distribution (4 min), together result in forks from one of these origins rarely passively replicating the other (in <1% of the population). Therefore, the efficiency of *ORI3* closely reflects its competence. In contrast, the close proximity of *ORI1* and *ORI2* (15 kb) results in forks from one origin occasionally passively replicating and therefore inactivating the other origin (in ~7% of the population). Consequently the efficiency of these origins is lower than their competence.

Sites of replication termination differ between members of the population (Figure 2B and Supplementary Figure S2A and B) as a consequence of each origin's competence and the stochasticity in origin activation time. It is important to note that this observation is based upon our assumption that there is a constant replication fork velocity. Fork velocity that depends strongly on the chromosomal coordinate would affect the distribution of replication termination sites. Similarly, variation in replication fork velocity between cells would increase the distribution range of termination sites. A result of variation in origin usage is that within the population each region of the chromosome is replicated by both leftward and rightward moving forks. The exceptions are those regions external (telomeric) to all three origins. This factor could potentially explain why nucleotide skew is only observed at telomeric sequences in *S. cerevisiae* (29).

Properties of T_{rep} curves

The T_{rep} output from the simulated replication of this virtual chromosome is shown in Figure 2C (solid line). To reproduce the smoothing applied to experimental datasets (15,16) we have applied a 4 kb smoothing window (dashed line). Without smoothing, peaks are sharp, with the gradient of the T_{rep} profile changing abruptly at the replication origins. Sharp peaks are a consequence of the fact that yeast replication origins have well-defined positions in the chromosome. In contrast, the smooth valleys result from the range of termination sites (see above), which in turn are a consequence of the stochasticity in activation times. Applying the smoothing window results in peaks being shifted in the direction of the shallower gradient (Figure 2C horizontal arrow) and a reduction in peak heights (vertical arrow), as a consequence of averaging. It is important to note that the only effect of asynchronous entry in to S phase on T_{rep} profiles is to shift them vertically; their shapes, including the gradients are unchanged. We have verified this observation by analysing the model mathematically and by direct simulation of asynchronous entry in to the cell cycle (Supplementary Figure S3). Under these conditions, the amount of vertical displacement is equal to half the amount of introduced asynchrony.

Despite the fact that all the origins have the same activation time, the lower competence of *ORI3* results in a lower peak (dashed arrow), reminiscent of an origin with a later activation time (15). To investigate this further we compared T_{rep} curves for virtual chromosomes in which we varied either the activation time (Figure 3A) or the competence (Figure 3B). Two independent sets of origin parameters can result in similar T_{rep} profiles (Figure 3C). Thus the peak height in a T_{rep} plot is influenced by both origin activation time and competence. In addition, the peak height for one origin is affected by the parameters and proximity of other origins. Therefore peak heights cannot be directly interpreted as origin activation time and T_{rep} experimental data alone cannot be used to determine all the underlying parameters uniquely. The observation that an experimental dataset does not determine all parameters uniquely is referred to as the identifiability problem (30).

We have assumed a constant fork velocity, and the replication model results in T_{rep} curve gradients that are not constant (Figure 2)—i.e. the gradient does not equate to the fork velocity as has been suggested previously (15). In fact the T_{rep} curve gradient is determined by the proportion of leftward and rightward moving forks at a given position (data not shown).

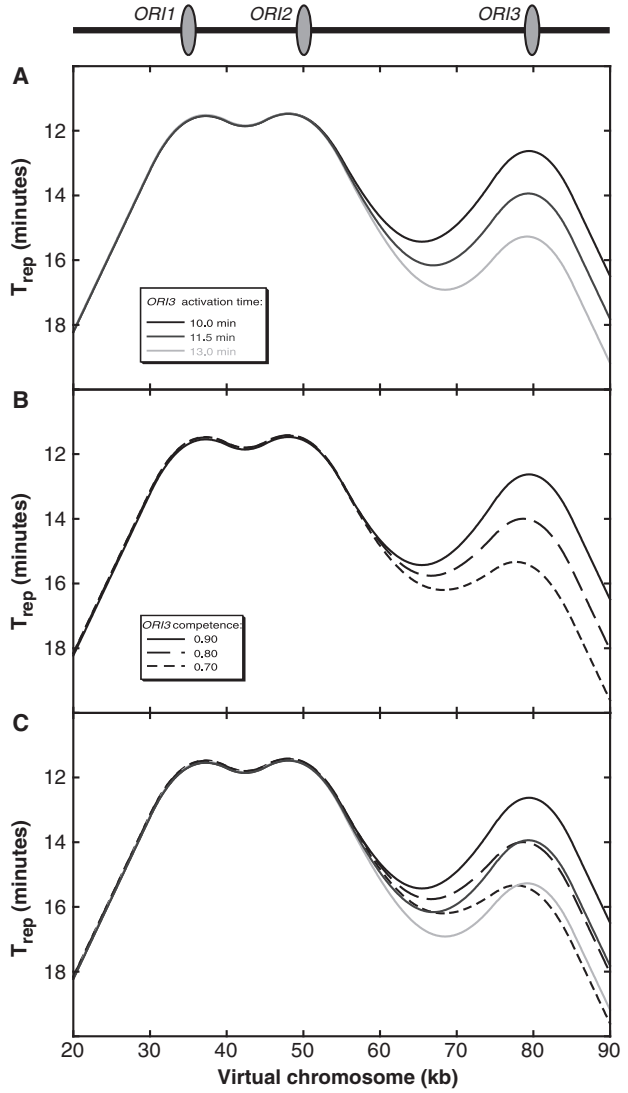

Figure 3. Mean T_{rep} plots illustrating the identifiability problem. Multiple parameter combinations can give rise to the same T_{rep} curve. The parameters for *ORI1* and *ORI2* remain the same throughout and are as described in Figure 2. (**A**) The consequence of varying *ORI3* activation time (10.0, 11.5 and 13.0 min.) on T_{rep} curves is shown (*ORI3* competence fixed at 0.9). (**B**) The consequence of varying *ORI3* competence (0.9, 0.8 and 0.7) on T_{rep} curves is shown (*ORI3* activation time fixed at 10 min). (**C**) Plots from (A) and (B) are superimposed to illustrate how similar profiles can arise from different parameter sets.

By simulating the replication of a virtual chromosome we have highlighted some of the complexities of the chromosome replication system. The biological parameters of the replication system cannot be determined directly from T_{rep} plot data because the shape of these plots is determined by the complex interplay between the parameters of many origins. This has led to the mis- or over-interpretation of T_{rep} plot data, as exemplified above. Therefore a mathematical model is invaluable for the interpretation of chromosome replication data.

Simulating the replication of a eukaryotic chromosome

We simulated the replication of *S. cerevisiae* chromosome VI. The replication of this chromosome has been systematically analysed by a number of complementary approaches. The positions of the origins have been determined by ARS assays, 2D-gels and various microarray studies (1). The results from these studies indicate that the replication of this chromosome is representative of other (less well studied) *S. cerevisiae* chromosomes, for example the origin density on this chromosome (~6 origins/100 kb) is comparable to the genome average (~5 origins/100 kb).

Table 1. Experimentally determined replication origin parameters

Replication origin	Position (kb)	Plasmid-based Competence	Average activation time (min)
ARS601/ARS602[a]	33	0.93	30
ARS603	69	0.95	25
ARS603.5	119	0.93	15
ARS604	128	0.58	22
ARS605	136	0.88	17.5
ARS606	168	0.92	15
ARS607	199	0.91	10
ARS608	217	0.90	20
ARS608.5/ARS609[a]	256	0.93	30

S. cerevisiae chromosome VI replication origin coordinates (1), competences based upon plasmid loss rates (32) and activation times based upon timed 2D-gels (33) are shown.
[a]These pairs of origins are considered to be a single origin site for modelling purposes because they are too close to be distinguished in genome-wide microarray datasets.

For our simulation, we considered nine origin sites (close pairs of origins are considered as a single origin site, since they are indistinguishable in microarray datasets; Table 1) covering ~230 kb of chromosome VI; each of these sites are particularly well studied (31–33) and this region excludes repetitive and telomeric sequences. Initially, we make the assumption that origin competence can be measured as the activity of an origin when it is isolated from other origins on a plasmid. The plasmid-based activities of chromosome VI origins have been experimentally determined (32–34) and are shown in Table 1. The mean chromosomal activation times for these origins have also been experimentally determined by 2D-gel analysis at 5 min time points through S phase (33). Although there has been no direct measure of the width of the activation time distribution (σ), timed 2D-gels indicate that this parameter is within the range of 5–20 min (33). Finally, replication fork velocity has been measured as ~1.4 kb/min in cells growing at 23°C (31).

Based upon these experimentally determined parameters (Table 1) we used our model to simulate the replication of chromosome VI and determine the mean T_{rep} across the chromosome. Initially, we considered all the origin activation times to have a probability distribution width (σ) of 7 min [consistent with timed 2D-gel experiments (33)]. The simulated data (Figure 4, black curve) closely resemble two independent experimental measurements of mean replication time. However it is important to note that a range of model parameters are likely to produce similar curves (see above). Nevertheless, this

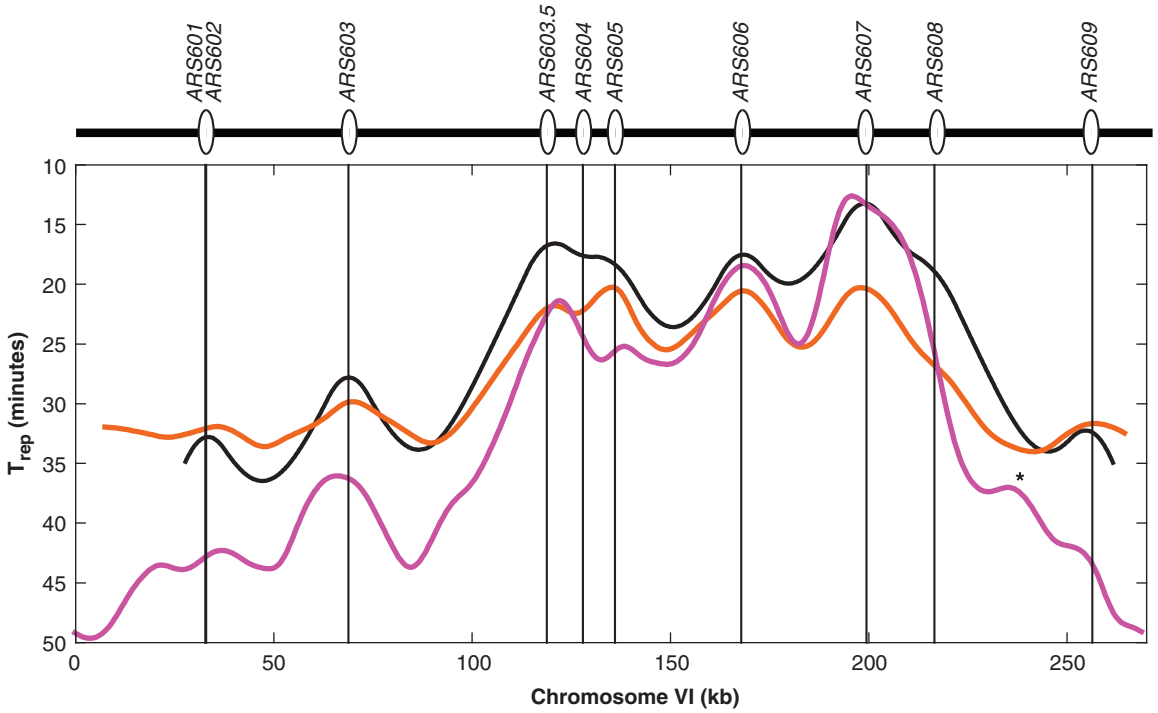

Figure 4. Simulation of *S. cerevisiae* chromosome VI replication. Mean T_{rep} from experimental datasets [purple line from (15); orange line from (16)] and from mathematical simulation using experimental parameters (black line). The location and names of the replication origins are shown in the cartoon (top) and by vertical lines on the plot. The peak in the purple curve at 236 kb (marked with asterisk) has since been shown to be an artefact and does not represent a replication origin (36).

result indicates that our model can correctly predict experimental data and offers a degree of validation for our model.

Next we examined how the width of origin activation time distribution (σ) influences replication time. Simulations were repeated using experimentally determined parameters (as above) but with a range of values for σ (from 0 to 15 min), again the same value of σ was used for each origin within each simulation. This range of values for σ did not greatly alter the resulting replication timing curves (Supplementary Figure S4A). However, it did alter the proportion of cells in which each origin activates (the origin efficiency). For example, the efficiency of *ARS607* falls slightly as σ increases but remains at ~85% when σ = 10 min. In contrast, *ARS608* efficiency falls rapidly and is only ~55% when σ = 10 min (Supplementary Figure S5A). By considering that origins activate stochastically within a window of time we can reproduce the observation that individual cells use different origin combinations to replicate a chromosome. Importantly this can be achieved without the necessity to assume a reduction in origin competence, thereby leaving dormant origins (e.g. *ARS604*) that can be envisaged to complete replication in the case of replicative stress.

Although stochastic origin activation results in varied origin usage it is noteworthy that no single value of σ (applied equally to all nine origin sites) results in predicted origin efficiencies in close agreement to experimental data. It is likely that each origin has a different value of σ, and it has recently been proposed that this value might increase in size the later the origin's mean activation time (35). Our model allows for σ to vary independently of the mean activation time (T), but to investigate this hypothesis we made σ a function of T ($\sigma_i = 0.5 \times T_i$). The resulting simulated replication times (Supplementary Figure S4B) are again in close agreement with the experimental data. This variable value of σ more closely predicts experimentally determined origin efficiencies than when a single value is applied to all origins (Supplementary Figure S5B). However, this simulation still overestimates origin efficiency for those origins that are rarely used (e.g. *ARS608*). Accurate experimental determination of cell-to-cell variability in origin activation time will be important for a better understanding of chromosome replication dynamics and to fully test whether or not σ and T are correlated.

Estimating system parameters from mean replication timing data

We used global optimization methods to estimate replication system parameters from experimental mean T_{rep} data (15). As demonstrated above, multiple combinations of origin activation time and competence can result in near identical mean replication timing curves (Figure 3). Therefore T_{rep} datasets cannot determine all parameters uniquely. To test our parameter estimation approach on T_{rep} data we assumed the location of the origins and their competence (Table 1) and used a version of the genetic algorithm (see 'Materials and methods' section) to

estimate values for origin activation times, fork velocity and a single value of σ for all origins (a total of 11 parameters). Estimated parameters were scored by the normalized sum of the squares of the differences between the simulated T_{rep} curve and the experimental curve. Parameter estimation was undertaken independently multiple times and each time the lowest scoring set of parameter values was retained. The genetic algorithm is a stochastic parameter estimation method which yields different results each time it runs; we assess the uncertainty in the estimation and highlight identifiability issues by running the algorithm multiple times. Comparison of these estimated parameters showed that some parameters are tightly determined by the data (for example the mean activation time of *ARS607*), whereas other parameters are poorly determined (for example the mean activation time of *ARS608*; Supplementary Figure S6). Therefore, mean replication timing data alone is not sufficient to define many of the parameters.

Estimating system parameters from time course data

Next we estimated replication system parameters from data that measure the fraction of DNA replicated at time points through S phase. These data have the potential to distinguish between different sets of parameters that result in near identical T_{rep} curves (Supplementary Figure S7) and may therefore help resolve the problem of identifiability, yielding tighter bounds on more parameters. We used recent experimental determination of the proportion of chromosome VI sequences that are replicated at various times during S phase (36) to estimate the values of all the replication system parameters for this chromosome using a genetic algorithm. This approach requires the microarray data and the location of the replication origins. Therefore, we can now consider a more extensive region of chromosome VI that includes *ARS600.3*/*ARS600.4* (as a pair) and *ARS610* (remaining origin sites do not fall within or close to unique sequence and therefore can not be analysed from microarray data). The requirement for confirmed origin locations does not currently allow us to apply this approach to all *S. cerevisiae* chromosomes, since to date only ~60% of origin sites have been experimentally confirmed (1).

We assumed origin locations (x_i) and estimated a total of 34 parameters—p_i, T_i and σ_i for each of the 11 origin sites, plus the replication fork velocity. Parameter estimation was undertaken using a genetic algorithm. An example of a fit is shown in Figure 5, where the curves represent the model-predicted percentage of replicated DNA and the points represent the raw experimental data. Repeating the parameter estimation multiple times and comparing the estimated parameters, indicates that the majority of estimated parameters are well defined by the data (the parameters from 614 highest scoring fits are given in Table S1). Our estimated origin activation times are within 5 min of the experimentally determined times for seven out of nine origins (the exceptions being *ARS605* and *ARS609*). The estimated values for σ vary from 3 min (*ARS607*) to 17 min (*ARS603*, *ARS608* and *ARS609*) and show a trend towards increased values of σ with later activation time (Supplementary Figure S8). This suggests that σ may be correlated with the mean activation time (T). Mean replication fork velocity (at 23°C) was estimated at 1.6 kb/min; this value agrees well with the independent experimental result of 1.4 kb/min (measured at 23°C) (31). Our estimated fork velocity is also in agreement with the modal fork velocity from Raghuraman *et al.* (2001) of between 1.5 and 2.0 kb/min (at 23°C) and from Yabuki *et al.* (2002) of 2.8 ± 1 kb/min (at 30°C). Furthermore, model estimates of the time to complete chromosome replication are consistent with experimental measurements (Supplementary Figure S9).

The predicted competence of each origin (see above) was compared to data from plasmid loss rate experiments (Supplementary Figure S10). In almost every case the estimated value was lower than that experimentally measured and this was most pronounced for the telomere proximal origins (e.g. *ARS609*). It seems that plasmid-based measurements of origin activity may overestimate the competence of an origin at its native chromosomal locus (this may particularly be the case for telomere proximal origins where the circular plasmid fails to mimic the chromosome end). Alternatively our model and/or parameter estimation may underestimate values for origin competence. However, since it is not clear that direct measurements of origin competence are possible we sought alternate methods for model validation.

Validating model predictions

We investigated the ability of our model to predict measurable system outputs. Based upon the parameters estimated from the Alvino *et al.* data, we made a number of predictions about the replication of chromosome VI. One output that has been independently determined experimentally is the percentage of cells in which a replication origin is active (origin efficiency). Two studies have determined origin efficiency, using different experimental approaches, for eight of the origins on chromosome VI (31,33). The approaches used are technically challenging and the errors in the reported values are likely to be high [estimated at ± 10% from comparison of *ARS603* and *ARS607* efficiencies reported by (31,37)]; which might explain the differences in values obtained for some origins between the two studies. Nevertheless, these experimental values provide an independent dataset to test the validity of our model and the estimated parameters.

Figure 6A shows experimentally determined origin efficiencies (31) and our model-predicted efficiencies (the mean value from 614 high scoring parameter estimation runs; Supplementary Table S2). For seven out of eight origins (*ARS601*/*2*, *ARS603*, *ARS603.5*, *ARS605*, *ARS606*, *ARS607* and *ARS608*) the model-predicted efficiencies are similar to the experimentally determined values (difference <10%; Figure 6A and Supplementary Figures S11 and S12). For *ARS609* the difference between the model prediction and experimentally determined efficiency is 20%. This origin (*ARS609*) lies close to the end of the chromosome and therefore there is only limited

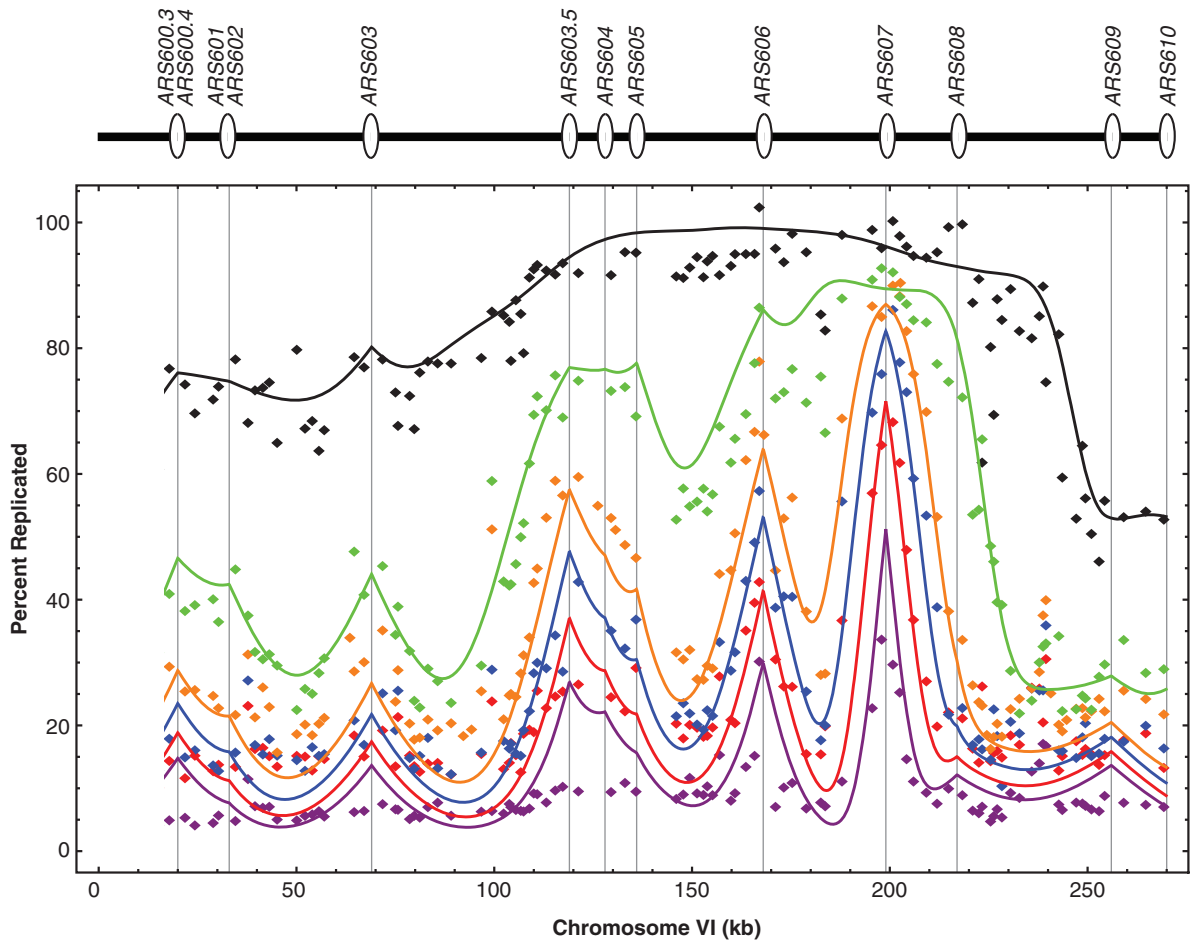

Figure 5. Percentage of replicated DNA for *S. cerevisiae* chromosome VI: model fit and experimental data. Markers are raw experimental data (36) and solid lines are model fits. The six curves (from top to bottom) correspond to the percentage replicated DNA at 10 (purple), 12.5 (red), 15 (blue), 17.5 (orange), 25 (green) and 40 (black) min. The location and names of the replication origins are shown in the cartoon (top) and by vertical lines on the plot.

microarray data available, potentially reducing the reliability of our parameter estimations. Nevertheless, our model-estimated parameters can quantitatively recapitulate the majority of experimentally observed differences in origin efficiency.

Finally, we made a number of further predictions about the replication of chromosome VI. These include the distribution of replication termination sites, the number of origins that are active on each chromosome within the population, the distribution of distances between active origins and the time taken to complete chromosome replication (Figure 6 and Supplementary Figure S13). We selected these outputs due to their biological relevance and the likelihood that they can be experimentally tested in the future, for example using single molecule techniques over a time course. Early single molecule (electron microscopy) studies on *S. cerevisiae* chromosome replication estimated the mean distance between active replication origins at 36 kb (38), in remarkable agreement with our model predictions (Figure 6D). These independent experimental validations of our model demonstrate its predictive power.

DISCUSSION

In this study we present a quantitative predictive stochastic model for eukaryotic chromosome replication. This model has allowed us to: (i) illustrate how different origin parameters contribute to differential origin efficiency (Supplementary Data 1); (ii) highlight the complexity in interpreting genome-wide chromosome replication data, specifically the issue of identifiability (Figure 3); (iii) use parameter estimation to determine the values of chromosome replication parameters (Supplementary Table S1); and (iv) demonstrate that differential origin efficiency can result from stochastic origin activation (rather than reduced origin competence) thereby leaving dormant origins with the ability to 'rescue' chromosome replication in the event of replicative stress. Finally, estimated parameters enable the prediction of origin efficiencies that are in close agreement with independent experimental data, allowing us to validate our model (Figure 6A).

We have focussed our work on *S. cerevisiae* chromosome VI. This is one of the smallest yeast chromosomes and the replication origins have been extensively

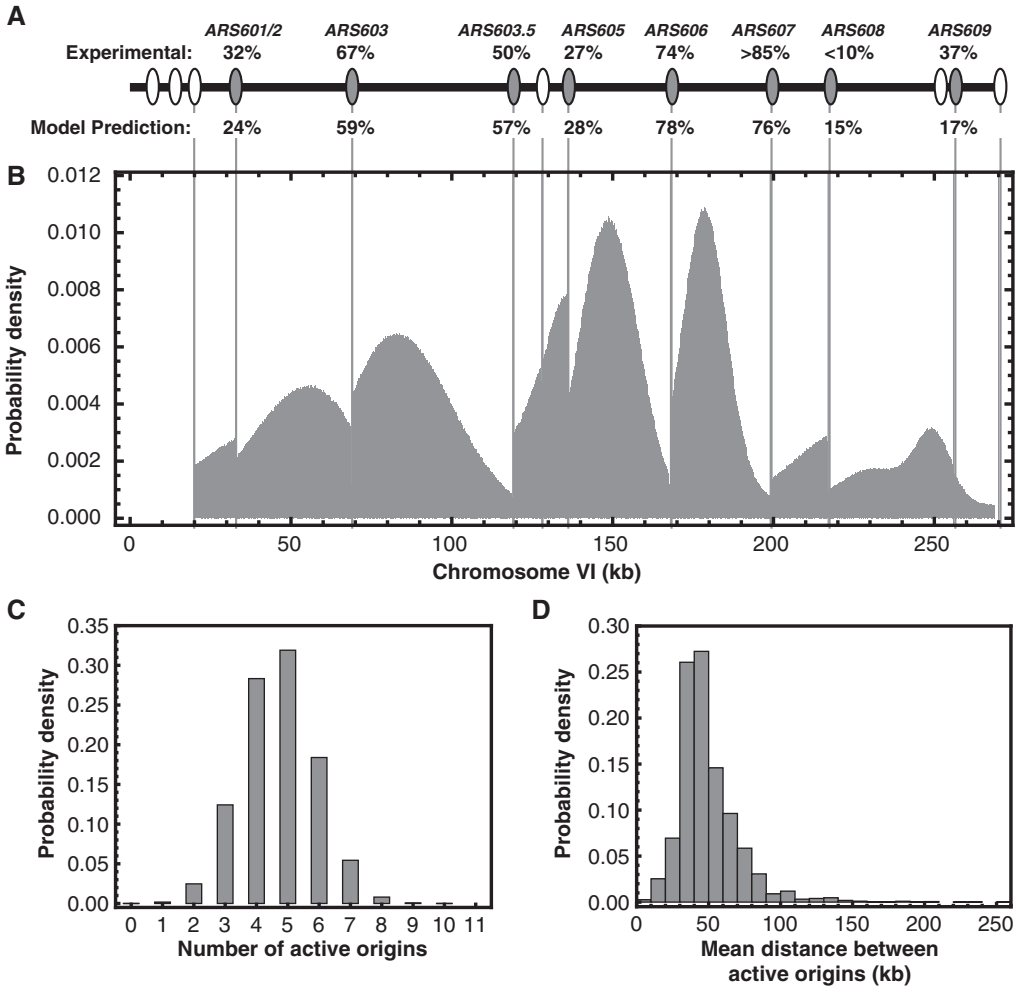

Figure 6. Prediction of *S. cerevisiae* chromosome VI replication dynamics. Parameters for 11 chromosome VI replication origin sites were determined by parameter estimation using replication time course data (36). Resulting parameters were used to simulate chromosome replication and predict chromosome replication dynamics. (**A**) Schematic of *S. cerevisiae* chromosome VI comparing experimentally determined (31) and simulated origin efficiencies (shaded and labelled ovals). (**B**) Simulated distribution of replication fork termination events. Telomeric termination events are not shown since they have a probability of 1. (Vertical lines illustrate the location of origin sites used in the simulation.) Notice the discontinuity of the distribution at the origin locations. (**C**) Distribution of the number of active chromosome VI origins. (**D**) Distribution of the mean distance between active origins (excluding those chromosomes where only one origin was active).

characterized. Preliminary analysis of other well-characterized chromosomes (II, III and X) gave comparable results (not shown). Despite the wealth of studies describing *S. cerevisiae* chromosome replication dynamics (15,16,35,36) the density of data points relative to origin density is occasionally low; for example there are only four data points between *ARS604* and *ARS605* and just one between *ARS603.5* and *ARS604* (36). The limited microarray data in this region could potentially reduce the reliability of our estimated parameters for these origins. Additional challenges are posed by the scarcity of data points towards chromosome ends. Perhaps the greatest challenge to modelling chromosome replication is in understanding cell-to-cell differences. Here we estimate how chromosome VI origins may differ in their activation time within a population. Single molecule (39) or single cell approaches (40) will be required to validate these predictions and will need to be able to distinguish between cell-cycle synchrony and origin activation synchrony.

Determining the molecular mechanisms that underlie differences in origin efficiency and time of activation is crucial for our understanding of chromosome replication. As a first step to investigating these mechanisms it is essential to correctly interpret current experimental data. Here we show that differences in both origin competence and origin activation time can contribute to differences in observed mean T_{rep}. Therefore it is not reliable to consider the T_{rep} of an origin sequence as a measure of origin activation time. This observation is crucial for correct determination of the factors influencing differences in origin efficiency, particularly when interpreting data from mutants that change the behaviour of replication origins (41–43).

The role of multiple origin parameters in determining origin usage highlights the importance of including all such parameters when modelling DNA replication. This study is the first to include the concept of origin competence in a model. Describing each origin with both a

competence and an activation time captures the essential biology of the system; that is a licensing stage (low CDK) and an activation stage (high CDK), which are mutually exclusive. Previous models of chromosome replication have either not considered origin competence (19,23) or have defined replication origins as licensed sites thereby avoiding the requirement for this parameter (20,22). Factors influencing the competence of an origin may include the affinity of the sequence for ORC (28) and other pre-RC components, the local chromatin structure (11) and the transcriptional environment (27). Recent experimental datasets are starting to define the location of replication origins in other eukaryotes; modelling of replication in these systems, based upon newly located replication origins, will benefit from consideration of origin competence.

A further distinction between this study and previous models is the way we have considered replication origin activation time. We have defined each origin to have a mean activation time, but each origin activates stochastically within a window around this mean time. Several previous models have considered origins to have a certain propensity to activate within a small time window (18,20,23,44). In these models origin activation is stochastic, however within the population those origins with a high activation propensity are likely to activate earlier than origins with a lower activation propensity; thus recreating (within a population) the observed ordered firing of replication origins (44). To reproduce experimental observations and overcome the random completion problem it is necessary for origin activation propensity to increase during S phase, and several biological explanations for this have been proposed and modelled (22,23). In our model, chromosome replication is predicted to be complete by 70 min (in >99% of molecules; Supplementary Figure S13D), which is significantly earlier than cell division. It is important to note that 'origin activation propensity' and 'origin competence' describe completely different origin properties. Origin competence is determined by the proficiency of pre-RC assembly, whereas origin activation propensity is determined by the proficiency of origin activation. Modelling of the molecular steps that determine these origin parameters will require more complex models and additional biochemical data.

Defining chromosome replication in generic terms, with encoding of the relevant dynamical features of each origin into distinct origin parameters, has allowed us to design a model that can be applied to a range of experimental systems from archaea to metazoans. In the future, we anticipate that the modelling of these varied systems will help elucidate the common mechanisms of replication origin regulation and genome replication.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Carolin Mueller and Drs Amy Upton, Shin-ichiro Hiraga and Anne Donaldson for critical reading of the manuscript.

FUNDING

The Leverhulme Trust; The University of Nottingham; The University of Aberdeen; and the Biotechnology and Biological Sciences Research Council (grant numbers BB/E023754/1, BB/G001596/1); CAN is a David Phillips Fellow. Funding for open access charge: Biotechnology and Biological Sciences Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Nieduszynski,C.A., Hiraga,S., Ak,P., Benham,C.J. and Donaldson,A.D. (2007) OriDB: a DNA replication origin database. *Nucleic Acids Res.*, **35**, D40–D46.
2. Huberman,J.A. and Riggs,A.D. (1968) On the mechanism of DNA replication in mammalian chromosomes. *J. Mol. Biol.*, **32**, 327–341.
3. Santocanale,C., Sharma,K. and Diffley,J.F. (1999) Activation of dormant origins of DNA replication in budding yeast. *Genes Dev.*, **13**, 2360–2364.
4. Woodward,A.M., Gohler,T., Luciani,M.G., Oehlmann,M., Ge,X., Gartner,A., Jackson,D.A. and Blow,J.J. (2006) Excess Mcm2-7 license dormant origins of replication that can be used under conditions of replicative stress. *J. Cell. Biol.*, **173**, 673–683.
5. Bell,S.P. and Dutta,A. (2002) DNA replication in eukaryotic cells. *Annu. Rev. Biochem.*, **71**, 333–374.
6. Diffley,J.F. (2004) Regulation of early events in chromosome replication. *Curr. Biol.*, **14**, R778–R786.
7. Stillman,B. (2005) Origin recognition and the chromosome cycle. *Febs Lett.*, **579**, 877–884.
8. Struhl,K., Stinchcomb,D.T., Scherer,S. and Davis,R.W. (1979) High-frequency transformation of yeast: autonomous replication of hybrid DNA molecules. *Proc. Natl Acad. Sci. USA*, **76**, 1035–1039.
9. Chen,Z., Speck,C., Wendel,P., Tang,C., Stillman,B. and Li,H. (2008) The architecture of the DNA replication origin recognition complex in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **105**, 10326–10331.
10. Lee,D.G. and Bell,S.P. (1997) Architecture of the yeast origin recognition complex bound to origins of DNA replication. *Mol. Cell. Biol.*, **17**, 7159–7168.
11. Nieduszynski,C.A., Knox,Y. and Donaldson,A.D. (2006) Genome-wide identification of replication origins in yeast by comparative genomics. *Genes Dev.*, **20**, 1874–1879.
12. Shor,E., Warren,C.L., Tietjen,J., Hou,Z., Muller,U., Alborelli,I., Gohard,F.H., Yemm,A.I., Borisov,L., Broach,J.R. *et al.* (2009) The origin recognition complex interacts with a subset of metabolic genes tightly linked to origins of replication. *PLoS Genet.*, **5**, e1000755.
13. Xu,W., Aparicio,J.G., Aparicio,O.M. and Tavare,S. (2006) Genome-wide mapping of ORC and Mcm2p binding sites on tiling arrays and identification of essential ARS consensus sequences in *S. cerevisiae*. *BMC Genomics*, **7**, 276.
14. Feng,W., Collingwood,D., Boeck,M.E., Fox,L.A., Alvino,G.M., Fangman,W.L., Raghuraman,M.K. and Brewer,B.J. (2006) Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. *Nat. Cell. Biol.*, **8**, 148–155.
15. Raghuraman,M.K., Winzeler,E.A., Collingwood,D., Hunt,S., Wodicka,L., Conway,A., Lockhart,D.J., Davis,R.W., Brewer,B.J. and Fangman,W.L. (2001) Replication dynamics of the yeast genome. *Science*, **294**, 115–121.

16. Yabuki,N., Terashima,H. and Kitada,K. (2002) Mapping of early firing origins on a replication profile of budding yeast. *Genes Cells*, **7**, 781–789.
17. Castrillo,J.I. and Oliver,S.G. (2004) Yeast as a touchstone in post-genomic research: strategies for integrative analysis in functional genomics. *J. Biochem. Mol. Biol.*, **37**, 93–106.
18. Goldar,A., Marsolier-Kergoat,M.C. and Hyrien,O. (2009) Universal temporal profile of replication origin activation in eukaryotes. *PLoS ONE*, **4**, e5899.
19. Spiesser,T.W., Klipp,E. and Barberis,M. (2009) A model for the spatiotemporal organization of DNA replication in *Saccharomyces cerevisiae*. *Mol. Genet. Genomics*, **282**, 25–35.
20. Blow,J.J. and Ge,X.Q. (2009) A model for DNA replication showing how dormant origins safeguard against replication fork failure. *EMBO Rep.*, **10**, 406–412.
21. Gauthier,M.G. and Bechhoefer,J. (2009) Control of DNA replication by anomalous reaction-diffusion kinetics. *Phys. Rev. Lett.*, **102**, 158104.
22. Goldar,A., Labit,H., Marheineke,K. and Hyrien,O. (2008) A dynamic stochastic model for DNA replication initiation in early embryos. *PLoS ONE*, **3**, e2919.
23. Lygeros,J., Koutroumpas,K., Dimopoulos,S., Legouras,I., Kouretas,P., Heichinger,C., Nurse,P. and Lygerou,Z. (2008) Stochastic hybrid modeling of DNA replication across a complete genome. *Proc. Natl Acad. Sci. USA*, **105**, 12295–12300.
24. Yang,S.C. and Bechhoefer,J. (2008) How Xenopus laevis embryos replicate reliably: investigating the random-completion problem. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **78**, 041917.
25. Czajkowsky,D.M., Liu,J., Hamlin,J.L. and Shao,Z. (2008) DNA combing reveals intrinsic temporal disorder in the replication of yeast chromosome VI. *J. Mol. Biol.*, **375**, 12–19.
26. Donato,J.J., Chung,S.C. and Tye,B.K. (2006) Genome-wide hierarchy of replication origin usage in *Saccharomyces cerevisiae*. *PLoS Genet.*, **2**, e141.
27. Nieduszynski,C.A., Blow,J.J. and Donaldson,A.D. (2005) The requirement of yeast replication origins for pre-replication complex proteins is modulated by transcription. *Nucleic Acids Res.*, **33**, 2410–2420.
28. Palacios DeBeer,M.A., Muller,U. and Fox,C.A. (2003) Differential DNA affinity specifies roles for the origin recognition complex in budding yeast heterochromatin. *Genes Dev.*, **17**, 1817–1822.
29. Gierlik,A., Kowalczuk,M., Mackiewicz,P., Dudek,M.R. and Cebrat,S. (2000) Is there replication-associated mutational pressure in the *Saccharomyces cerevisiae* genome? *J. Theor. Biol.*, **202**, 305–314.
30. Wilkinson,D.J. (2009) Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.*, **10**, 122–133.
31. Friedman,K.L., Brewer,B.J. and Fangman,W.L. (1997) Replication profile of *Saccharomyces cerevisiae* chromosome VI. *Genes to Cells*, **2**, 667–678.
32. Shirahige,K., Iwasaki,T., Rashid,M.B., Ogasawara,N. and Yoshikawa,H. (1993) Location and characterization of autonomously replicating sequences from chromosome VI of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **13**, 5043–5056.
33. Yamashita,M., Hori,Y., Shinomiya,T., Obuse,C., Tsurimoto,T., Yoshikawa,H. and Shirahige,K. (1997) The efficiency and timing of initiation of replication of multiple replicons of *Saccharomyces cerevisiae* chromosome VI. *Genes Cells*, **2**, 655–666.
34. Chang,F., Theis,J.F., Miller,J., Nieduszynski,C.A., Newlon,C.S. and Weinreich,M. (2008) Analysis of chromosome III replicators reveals an unusual structure for the ARS318 silencer origin and a conserved WTW sequence within the origin recognition complex binding site. *Mol. Cell. Biol.*, **28**, 5071–5081.
35. McCune,H.J., Danielson,L.S., Alvino,G.M., Collingwood,D., Delrow,J.J., Fangman,W.L., Brewer,B.J. and Raghuraman,M.K. (2008) The temporal program of chromosome replication: genomewide replication in clb5{Delta} *Saccharomyces cerevisiae*. *Genetics*, **180**, 1833–1847.
36. Alvino,G.M., Collingwood,D., Murphy,J.M., Delrow,J., Brewer,B.J. and Raghuraman,M.K. (2007) Replication in hydroxyurea: it's a matter of time. *Mol. Cell. Biol.*, **27**, 6396–6406.
37. Donaldson,A.D., Raghuraman,M.K., Friedman,K.L., Cross,F.R., Brewer,B.J. and Fangman,W.L. (1998) CLB5-dependent activation of late replication origins in *S. cerevisiae*. *Mol. Cell*, **2**, 173–182.
38. Newlon,C.S. and Burke,W.G. (1980) Replication of small chromosomal DNAs in yeast. In Alberts,B.A. and Stusser,F.C. (eds), *Mechanistic studies of DNA replication and genetic recombination*. Academic Press, New York, pp. 399–409.
39. Tuduri,S., Tourriere,H. and Pasero,P. (2010) Defining replication origin efficiency using DNA fiber assays. *Chromosome Res.*, **18**, 91–102.
40. Kitamura,E., Blow,J.J. and Tanaka,T.U. (2006) Live-cell imaging reveals replication of individual replicons in eukaryotic replication factories. *Cell*, **125**, 1297–1308.
41. Cosgrove,A.J., Nieduszynski,C.A. and Donaldson,A.D. (2002) Ku complex controls the replication time of DNA in telomere regions. *Genes Dev.*, **16**, 2485–2490.
42. Knott,S.R.V., Viggiani,C.J., Tavare,S. and Aparicio,O.M. (2009) Genome-wide replication profiles indicate an expansive role for Rpd3L in regulating replication initiation timing or efficiency, and reveal genomic loci of Rpd3 function in *Saccharomyces cerevisiae*. *Genes Dev.*, **23**, 1077–1090.
43. Wu,P.Y. and Nurse,P. (2009) Establishing the program of origin firing during S phase in fission yeast. *Cell*, **136**, 852–864.
44. Rhind,N. (2006) DNA replication timing: random thoughts about origin firing. *Nat. Cell. Biol.*, **8**, 1313–1316.